Supplementary Material for: Hierarchical Visual-Textual Graph for Temporal Activity Localization via Language

Shaoxiang Chen and Yu-Gang Jiang[⋆]

Shanghai Key Lab of Intelligent Information Processing,
School of Computer Science, Fudan University
{sxchen13, ygj}@fudan.edu.cn

1 More on Motivation of the C^3G Module

The temporal convolution layers have captured the relationship between the visual and textual modalities to some extent. However, different feature channels in convolutional networks carry distinct information (this is a widely accepted concept for CNN), and some channels should be more important for computing visual-textual relevance. Motivated by this, we utilize the Cross-Channel Communication Graph (C^3G) to encourage information passing across feature channels. In the C^3G, information is communicated between all pairs of channels (including visual and textual, since they're concatenated) so that each channel's importance is computed after taking into consideration information from both modalities in all the channels.

2 Effectiveness of the Loss Terms

Table 1. Performances of our model variants with different loss configurations.

#	Loss_b	Loss_r	Loss_a	Loss_n	IoU=0.3	IoU=0.5	IoU=0.7
0	✓	✗	✗	✗	41.91	27.28	8.82
1	✓	✓	✗	✗	56.29	42.45	19.84
2	✓	✓	✓	✗	58.82	44.97	21.77
3	✓	✓	✓	✓	61.37	47.27	23.30

We also study the effectiveness of each loss term in Eq (19), and the experimental results are shown in Table 1. The Boundary Loss (Loss_b) is non-optional because it supervises the final prediction layer. The Relevance Loss (Loss_r) is responsible for supervising the goal of our sentence localizer, which is computing the position-wise visual-textual relevance scores. Thus adding Loss_r leads to the largest performance gain. The Normalization Loss (Loss_n) and Alignment Loss (Loss_a) are mainly for normalizing and aligning the visual features with the textual features, and as we can observe from Table 1, they are both effective.

[⋆] Corresponding author.